Statistical Techniques in Robotics (16-831, S21) Lecture #08 (Monday, March 1)

Online Convex Optimization (OMD, Duality, Regret)

Lecturer: Kris Kitani *Scribes: Haowen Shi, Fan Jia*

1 Review

Previously we studied Follow The Leader (FTL) and Follow The Regularized Leader (FTRL) algorithms in the Online Convex Optimization (OCO) [2] category.

1.1 Follow The Leader with Quadratic Loss

Algorithm 1 FTL with Quadratic Loss

function FOLLOW THE LEADER (QUADRATIC LOSS)
  $f^{(0)} \leftarrow 0$
  for $t = 1, 2, \cdots, T$ **do**
    $\boldsymbol{w}^{(t)} = \arg\min_{\boldsymbol{w} \in S} \Sigma_{i=1}^{t-1} f^{(i)}(\boldsymbol{w})$
    RECEIVE $(f^{(t)} = \frac{1}{2}||\boldsymbol{w} - \boldsymbol{z}^{(t)}||_2^2)$ ▷ Receive loss function
  end for
end function

We proved the regret bound is:

$$\text{Regret} \leq 4L^2 \left(\log(T) + 1\right)$$

This is no-regret since the average regret grows sub-linearly in T. However, this only holds for quadratic loss functions, not for all convex functions.

1.2 Follow The Leader with Linear Loss

Algorithm 2 Follow the leader algorithm with linear loss

function FOLLOW THE LEADER (LINEAR LOSS)
  for $t = 1, 2, \cdots, T$ **do**
    $\boldsymbol{w}^{(t)} = \arg\min_{\boldsymbol{w} \in S} \left(\Sigma_{i=1}^{t-1} \boldsymbol{z}^{(i)}\right) \cdot \boldsymbol{w}$
    RECEIVE $(f^{(t)}(\boldsymbol{w}) = \boldsymbol{z}^{(t)} \cdot \boldsymbol{w})$ ▷ Receive loss function
  end for
end function

We observed that the regret of FTL with linear loss relative to an expert u that chooses $w = 0$ is:

$$\text{Regret}(u) = T - 1 - 0 = T - 1 = O(T)$$

The regret grows linearly because the $\boldsymbol{w}$ value flip-flops back and forth and make the algorithm unstable. To resolve this issue, we introduced regularization.

1.3 Follow The Regularized Leader

Algorithm 3 Follow The Regularized Leader

function FTRL(Convex Set $\mathcal{S}$)
  for $t = 1, 2, \cdots, T$ **do**
    $\boldsymbol{w}^{(t)} = \arg\min_{\boldsymbol{w} \in S} \Sigma_{i=1}^{t-1} f^{(i)}(\boldsymbol{w}) + \psi(\boldsymbol{w})$ ▷ Added $\psi(\boldsymbol{w})$ to the prediction rule
    Receive $(f^{(t)} : \mathcal{S} \to \mathbb{R})$
  end for
end function

We proved the regret bound for FTRL is:

$$\begin{aligned} R(\boldsymbol{u}) &\triangleq \sum_{t=1}^{T} \left[f^{(t)}(\boldsymbol{w}^{(t)}) - f^{(t)}(\boldsymbol{u}) \right] \\ &\leq \left[\psi(\boldsymbol{u}) - \psi(\boldsymbol{w}^{(1)}) \right] + \sum_{t=1}^{T} \left[f^{(t)}(\boldsymbol{w}^{(t)}) - f^{(t)}(\boldsymbol{w}^{(t+1)}) \right] \end{aligned}$$

Algorithm 4 FTRL with Linear Loss And Quadratic Regulation

function FTRL(Convex Set $\mathcal{S}$)
  for $t = 1, 2, \cdots, T$ **do**
    $\boldsymbol{w}^{(t)} = \arg\min_{\boldsymbol{w} \in S} \Sigma_{i=1}^{t-1} \left(\frac{1}{2\eta} ||\boldsymbol{w}||_2^2 + \boldsymbol{w} \cdot \sum_{i=1}^{t-1} z^{(i)} \right)$ ▷ Prediction rule
    Receive $(f^{(t)} : \boldsymbol{w} \cdot z^{(t)})$
  end for
end function

We proved the regret bound for FTRL with a Euclidean regularizer and linear loss function is:

$$R^{(T)}(\boldsymbol{u}) \leq BL\sqrt{2T}$$

Where $L = \max_{\mathcal{Z}} ||\mathcal{Z}||_2$ and $B = \max_{\boldsymbol{u} \in \mathcal{S}} ||\boldsymbol{u}||_2$.

Online Mirror Descent (OMD) can be considered as a special case of FTRL with a linear loss and a convex regularizer. In this lecture, we will focus on OMD.

2 Online Mirror Descent

Online Mirror Descent (OMD) [1] can be considered as a special case of FTRL. Fig. 1 shows the relationship between FTL, FTRL and OMD. FTRL adds a regularizer by forcing the zero-step loss to be a non-zero regularization function $\psi(\boldsymbol{w})$. In particular, if a FTRL problem has a convex regularization $\psi(\boldsymbol{w})$ and a linear loss function, it is called OMD.

With a linear loss, we can write the loss function as a dot product (linear combination) of the loss function parameter $\boldsymbol{z}$ and primal space parameter $\boldsymbol{w}$, as shown in Algorithm 5.

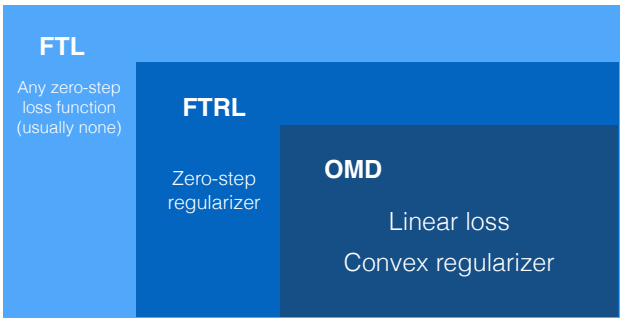

Figure 1: OMD can be seen as FTRL with convex regularization and linear loss

Algorithm 5 Follow The Regularized Leader - Linear Loss

function FTRL-LinLoss(Convex Set $\mathcal{S}$)
  for $t = 1, 2, \cdots, T$ **do**
    $\boldsymbol{w}^{(t)} = \arg\min_{\boldsymbol{w} \in S} \Sigma_{i=1}^{t-1} \langle \boldsymbol{w}, \boldsymbol{z}^{(i)} \rangle + \psi(\boldsymbol{w})$ ▷ Added $\psi(\boldsymbol{w})$ to the prediction rule
    Receive $(f^{(t)} : \mathcal{S} \to \mathbb{R})$
  end for
end function

We generalize `FTRL-LinLoss` to derive OMD. First, we define some notations:

Denote sum of $\boldsymbol{z}$:

$$\boldsymbol{z}^{(1:t)} = \sum_{i=1}^{t} \boldsymbol{z}^{(i)} \tag{1}$$

Define parameter of the dual space $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}^{(t+1)} \triangleq -\boldsymbol{z}^{(1:t)} \tag{2}$$

So we have the following update in each iteration for $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \boldsymbol{z}^{(t)} \tag{3}$$

Then, because the loss function $f^{(i)}$ is linear, we can generalize the FTRL prediction step as follows:

$$\begin{aligned}
\boldsymbol{w}^{(t+1)} &= \arg\min_{\boldsymbol{w}} \sum_{i=1}^{t} f^{(i)}(\boldsymbol{w}) + \psi(\boldsymbol{w}) && \text{FTRL prediction step} \\
&= \arg\min_{\boldsymbol{w}} \langle \boldsymbol{w}, \boldsymbol{z}^{(1:t)} \rangle + \psi(\boldsymbol{w}) && \text{Linear loss function} \\
&= \arg\max_{\boldsymbol{w}} \langle \boldsymbol{w}, -\boldsymbol{z}^{(1:t)} \rangle - \psi(\boldsymbol{w}) && \text{Convert to maximization} \\
&= \arg\max_{\boldsymbol{w}} \langle \boldsymbol{w}, \boldsymbol{\theta}^{(t+1)} \rangle - \psi(\boldsymbol{w}) && \text{Substitute (2)} \\
&= g(\boldsymbol{\theta}^{(t+1)}) && g\text{: mirror/linking function}
\end{aligned}$$

The function $g : \boldsymbol{\theta} \to \boldsymbol{w}$ is called the mirror/linking function because it takes parameters from the dual space ($\boldsymbol{\theta}$) and maps to the primal space ($\boldsymbol{w}$). This function is where the "mirror" name in the OMD came from. With the definitions introduced above, we can change some notations (without changing the algorithm itself) and re-write Algorithm 5 as:

Algorithm 6 Online Mirror Descent

function OMD(Convex Set $\mathcal{S}$, $g : \mathbb{R}^D \to \mathcal{S}$)
  for $t = 1, 2, \cdots, T$ **do**
    Receive ($f^{(t)} : \mathcal{S} \to \mathbb{R}$)
    $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta \boldsymbol{z}^{(t)}$, $\boldsymbol{z} \in \partial f^{(t)}(\boldsymbol{w}^{(t)})$ ▷ Dual parameter update
    $\boldsymbol{w}^{(t+1)} = g(\boldsymbol{\theta}^{(t+1)})$ ▷ Primal parameter update (or mirror projection)
  end for
end function

The regret bound of OMD (discussed in section 4) shares the same structure with FTRL because at the core they are the same algorithm. With OMD we are doing optimization in the dual space and map back to the primal space to get the updated parameter that we are optimizing (as illustrated in Fig. 2). The benefit of introducing dual space is primarily for theoretical purposes. By looking at the optimization in a different space, we can better understand optimization in the primal space and make more informed changes and improvements.

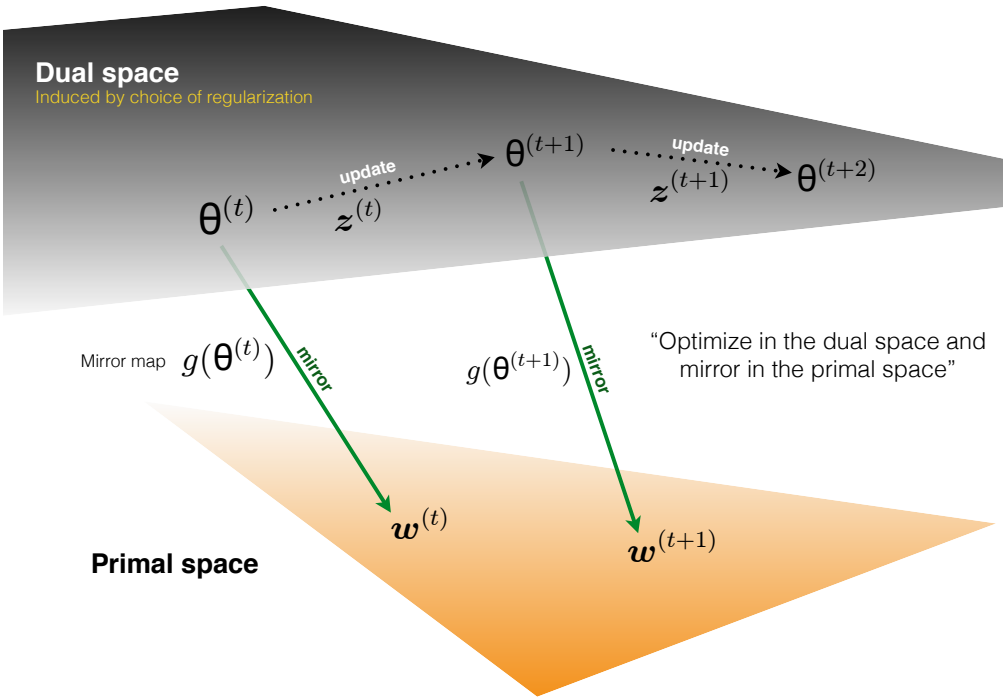

Figure 2: Illustration of online mirror descent in the dual space

A different regularization function ψ leads to a different mirror function g. The regularization should be chosen such that the mirroring can take better advantage of the "geometry" of the problem.

3 Duality

In this section we introduce mathematical tools that help us analyze OMD.

3.1 Convex Conjugate

A *convex conjugate* function is also called a Fenchel dual / a Fenchel conjugate / a Fenchel transform [4]. Here we focus on conjugate functions assuming a smooth convex function. The generalized version of this process for non-smooth and non-convex functions is called a Legendre transformation. The definition for a convex conjugate function is:

$$\psi^*(\theta) = \max_w(\langle \theta, w \rangle - \psi(w)) \tag{4}$$

For better understanding, the geometric interpretation of convex conjugate is included in the appendix section 5.1. Given (4), we can derive some useful properties used to analyze the regret of OMD:

Derivative of convex conjugate:

$$\nabla_{\boldsymbol{\theta}} \psi^*(\boldsymbol{\theta}) = \frac{\partial \psi^*(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \boldsymbol{w}^* = \arg\max_w (\langle \boldsymbol{\theta}, \boldsymbol{w} \rangle - \psi(\boldsymbol{w})) \tag{5}$$

$$\nabla_{\boldsymbol{w}} \psi(\boldsymbol{w}) = \left. \frac{\partial \psi(\boldsymbol{w})}{\partial \boldsymbol{w}} \right|_{\boldsymbol{w}=\boldsymbol{w}^*} = \boldsymbol{\theta} \tag{6}$$

Fenchel-Young inequality:

$$\psi^*(\boldsymbol{\theta}) \geq (\langle \boldsymbol{w}, \boldsymbol{\theta} \rangle - \psi(\boldsymbol{w})) \tag{7}$$

The proof for (5) is in section 5.2. The proof for (7) is trivial because any w other than the optimal w^* will produce less distance ψ than maximum ψ^* according to the maximizer definition (4).

3.2 Bregman Divergence

Bregman divergence is the approximation error of the first order approximation of $\psi(\boldsymbol{w})$ at $\boldsymbol{u}$.

$$D_\psi(\boldsymbol{w}||\boldsymbol{u}) = \psi(\boldsymbol{w}) - \psi(\boldsymbol{u}) - \nabla\psi(\boldsymbol{u})^\top(\boldsymbol{w} - \boldsymbol{u})$$

The geometric interpretation of Bregman divergence is included in the appendix section 5.3.

4 OMD Regret Analysis

We analysis the regret bound of the OMD algorithm in this section. The regret is:

$$R(\boldsymbol{u}) = \sum_{t=1}^{T} \boldsymbol{w}^{(t)} \cdot \boldsymbol{z}^{(t)} - \boldsymbol{u} \cdot \boldsymbol{z}^{(t)}$$

where $\boldsymbol{z}$ is the derivative of the loss function. The following regret bound holds for any OMD algorithm:

$$R(\boldsymbol{u}) \leq \psi(\boldsymbol{u}) - \psi(\boldsymbol{w}^{(1)}) + \sum_{t=1}^{T} D_{\psi^*}(-\boldsymbol{z}^{(1:t)}|| - \boldsymbol{z}^{(1:t-1)})$$

The regret is upper bounded by the sum of the difference in the regularization function plus the sum of all Bregman divergence under the convex conjugate of the regularization function. The proof is as follows:

Proof. The loss of an arbitrary vector $\boldsymbol{u}$:

$$\psi(\boldsymbol{u}) + \sum_{t=1}^{T} \boldsymbol{u} \cdot \boldsymbol{z}^{(t)} = \psi(\boldsymbol{u}) - \boldsymbol{u} \cdot \boldsymbol{\theta}^{T+1}$$

where $\boldsymbol{\theta}^{T+1}$ is the sum of gradients till to time step T.

Recall Fenchel-Young Inequality:

$$\psi^*(\boldsymbol{\theta}) \geq (\langle \boldsymbol{w}, \boldsymbol{\theta} \rangle - \psi(\boldsymbol{w}))$$

Apply Fenchel-Young Inequality to the right hand side:

$$\psi(\boldsymbol{u}) - \boldsymbol{u} \cdot \boldsymbol{\theta}^{T+1} \geq -\psi^*(\boldsymbol{\theta}^{(T+1)})$$

Then we expand the right hand side with telescoping. A telescoping series is a series whose partial sums eventually only have a fixed number of terms after cancellation.

Convert to telescoping series by adding and subtracting the complex conjugate at time step 1:

$$-\psi^*(\boldsymbol{\theta}^{(T+1)}) = -\psi^*(\boldsymbol{\theta}^{(T+1)}) - \psi^*(\boldsymbol{\theta}^{(1)}) + \psi^*(\boldsymbol{\theta}^{(1)})$$

Sum over shifted time steps:

$$\begin{aligned} -\psi^*(\boldsymbol{\theta}^{(T+1)}) &= -\psi^*(\boldsymbol{\theta}^{(T+1)}) - \psi^*(\boldsymbol{\theta}^{(T)}) + \psi^*(\boldsymbol{\theta}^{(T)}) - \cdots - \psi^*(\boldsymbol{\theta}^{(1)}) + \psi^*(\boldsymbol{\theta}^{(1)}) \\ &= -\psi^*(\boldsymbol{\theta}^{(1)}) - \sum_{t=1}^{T} (\psi^*(\boldsymbol{\theta}^{(t+1)}) - \psi^*(\boldsymbol{\theta}^{(t)})) \end{aligned}$$

Plug in definition of Bregman divergence:

$$-\psi^*(\boldsymbol{\theta}^{(T+1)}) = -\psi^*(\boldsymbol{\theta}^{(1)}) - \sum_{t=1}^{T} \left(\nabla \psi^*(\boldsymbol{\theta}^{(t)}) \cdot (\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}) + D_{\psi^*}(\boldsymbol{\theta}^{(t+1)} || \boldsymbol{\theta}^{(t)}) \right)$$

The first term can be written as $\psi(\boldsymbol{w}^{(1)})$ since:

$$\begin{aligned} \psi^*(\boldsymbol{\theta}^{(1)}) &= \psi^*(\boldsymbol{z}^{(0)}) && \text{from definition } \boldsymbol{\theta}^{(t+1)} \triangleq -\boldsymbol{z}^{(1:t)} \\ &= \psi^*(\mathbf{0}) && \text{From definition of conjugate} \\ &= \max_{\boldsymbol{w}} (\langle \boldsymbol{w}, \mathbf{0} \rangle - \psi(\boldsymbol{w})) && \text{Compute dot product} \\ &= \max_{\boldsymbol{w}} (\mathbf{0} - \psi(\boldsymbol{w})) && \text{Sign flip max-min conversion} \\ &= -\min_{\boldsymbol{w}} \psi(\boldsymbol{w}) && \text{Plug in minimizer of cumulative loss at time 1} \\ &= -\psi(\boldsymbol{w}^{(1)}) && \text{Minimizer is the primal iterate at step 1} \\ & && \text{('the one-step look ahead cheater')} \end{aligned}$$

Therefore, we have:

$$-\psi^*(\boldsymbol{\theta}^{(T+1)}) = \psi(\boldsymbol{w}^{(1)}) - \sum_{t=1}^{T} \Big(\nabla\psi^*(\boldsymbol{\theta}^{(t)}) \cdot (\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}) + D_{\psi^*}(\boldsymbol{\theta}^{(t+1)} || \boldsymbol{\theta}^{(t)}) \Big)$$

Plug in definition of dual parameter $\boldsymbol{\theta}^{(t+1)} \triangleq -\boldsymbol{z}^{(1:t)}$:

$$-\psi^*(\boldsymbol{\theta}^{(T+1)}) = \psi(\boldsymbol{w}^{(1)}) - \sum_{t=1}^{T} \Big(\nabla\psi^*(-\boldsymbol{z}^{(1:t-1)}) \cdot (-\boldsymbol{z}^{(1:t)} + \boldsymbol{z}^{(1:t-1)}) + D_{\psi^*}(-\boldsymbol{z}^{(1:t)} || -\boldsymbol{z}^{(1:t-1)}) \Big)$$

Since $\nabla_\theta \psi^*(\boldsymbol{\theta}) = \frac{\partial \psi^*(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \boldsymbol{w}^*$ and $-\boldsymbol{z}^{(1:t)} + \boldsymbol{z}^{(1:t-1)} = -\boldsymbol{z}^{(t)}$:

$$-\psi^*(\boldsymbol{\theta}^{(T+1)}) = \psi(\boldsymbol{w}^{(1)}) - \sum_{t=1}^{T} \Big(\langle \boldsymbol{w}^{(t)}, -\boldsymbol{z}^{(t)} \rangle + D_{\psi^*}(-\boldsymbol{z}^{(1:t)} || -\boldsymbol{z}^{(1:t-1)}) \Big)$$

Move negative out of sum, we can get:

$$-\psi^*(\boldsymbol{\theta}^{(T+1)}) = \psi(\boldsymbol{w}^{(1)}) + \sum_{t=1}^{T} \Big(\langle \boldsymbol{w}^{(t)}, \boldsymbol{z}^{(t)} \rangle - D_{\psi^*}(-\boldsymbol{z}^{(1:t)} || -\boldsymbol{z}^{(1:t-1)}) \Big)$$

Therefore:

$$\psi^*(-\boldsymbol{z}^{(1:T)}) = -\psi(\boldsymbol{w}^{(1)}) - \sum_{t=1}^{T} \Big(\langle \boldsymbol{w}^{(t)}, \boldsymbol{z}^{(t)} \rangle - D_{\psi^*}(-\boldsymbol{z}^{(1:t)} || -\boldsymbol{z}^{(1:t-1)}) \Big)$$

Recall the Fenchel-Young Inequality, we can have:

$$\langle \boldsymbol{u}, -\boldsymbol{z}^{(1:T)} \rangle - \psi(\boldsymbol{u}) \leq \psi^*(-\boldsymbol{z}^{(1:T)})$$

$$\langle \boldsymbol{u}, -\boldsymbol{z}^{(1:T)} \rangle - \psi(\boldsymbol{u}) \leq -\psi(\boldsymbol{w}^{(1)}) - \sum_{t=1}^{T} \Big(\langle \boldsymbol{w}^{(t)}, \boldsymbol{z}^{(t)} \rangle - D_{\psi^*}(-\boldsymbol{z}^{(1:t)} || -\boldsymbol{z}^{(1:t-1)}) \Big)$$

$$\langle \boldsymbol{u}, -\boldsymbol{z}^{(1:T)} \rangle - \psi(\boldsymbol{u}) \leq -\psi(\boldsymbol{w}^{(1)}) - \sum_{t=1}^{T} \langle \boldsymbol{w}^{(t)}, \boldsymbol{z}^{(t)} \rangle + \sum_{t=1}^{T} D_{\psi^*}(-\boldsymbol{z}^{(1:t)} || -\boldsymbol{z}^{(1:t-1)})$$

Rearranging this equation, we can have:

$$\sum_{t=1}^{T} \langle \boldsymbol{w}^{(t)}, \boldsymbol{z}^{(t)} \rangle - \langle \boldsymbol{u}, \boldsymbol{z}^{(1:T)} \rangle \leq \psi(\boldsymbol{u}) - \psi(\boldsymbol{w}^{(1)}) + \sum_{t=1}^{T} D_{\psi^*}(-\boldsymbol{z}^{(1:t)} || -\boldsymbol{z}^{(1:t-1)})$$

Therefore, our general OMD regret bound is:

$$\begin{aligned} R(\boldsymbol{u}) &= \sum_{t=1}^{T} \boldsymbol{w}^{(t)} \cdot \boldsymbol{z}^{(t)} - \boldsymbol{u} \cdot \boldsymbol{z}^{(t)} \\ &\leq \psi(\boldsymbol{u}) - \psi(\boldsymbol{w}^{(1)}) + \sum_{t=1}^{T} D_{\psi^*}(-\boldsymbol{z}^{(1:t)} || -\boldsymbol{z}^{(1:t-1)}) \end{aligned}$$

□

References

[1] G. Lan, A. Nemirovski, and A. Shapiro. Validation analysis of mirror descent stochastic approximation method. *Mathematical programming*, 134(2):425–458, 2012.

[2] S. Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and trends in Machine Learning*, 4(2):107–194, 2011.

[3] Wikipedia contributors. Bregman divergence — Wikipedia, the free encyclopedia, 2021. [Online; accessed 03-Mar-2021].

[4] Wikipedia contributors. Convex conjugate — Wikipedia, the free encyclopedia, 2021. [Online; accessed 03-Mar-2021].

5 Appendix (Covered in Lecture)

5.1 Geometric interpretation of convex conjugate

5.1.1 Intercept-slope parameterization

There is more than one way to parameterize a function. One method that we are familiar with the most is the function-value parameterization, where a single output value is provided for each input. Another way we can parameterize a function is called a intercept-slope parameterization.

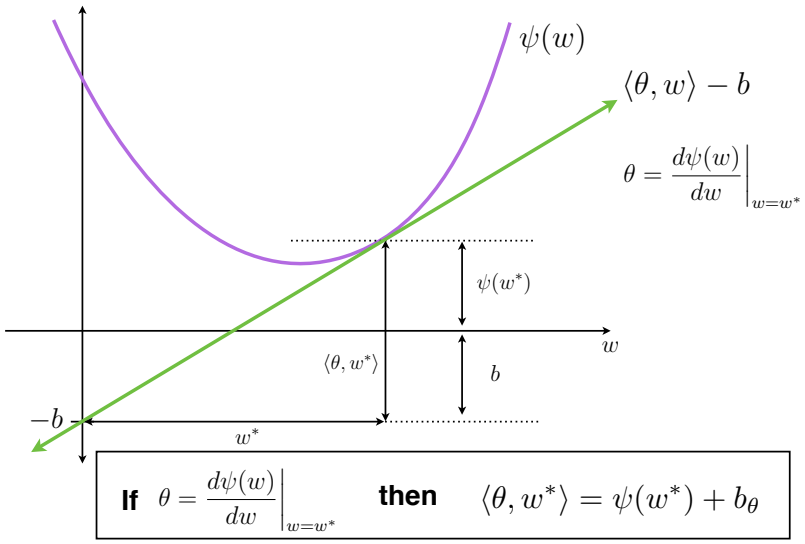

Figure 3: Geometry of tangents

As shown in Fig. 3, we have a function $\psi(w)$, and we define θ as the derivative evaluated at $w = w^*$. Then, the tangent line can be written as $\langle \theta, w \rangle - b$. So, the vertical distance ("rise") between $-b$ and the tangent point is $\langle \theta, w \rangle$. Additionally, the vertical distance can also be written as $\psi(w^*) + b$, because of the geometry shown in the figure. So, we have:

$$\langle \theta, w^* \rangle = \psi(w^*) + b(\theta) \tag{8}$$

Where $b(\theta)$ depends on the θ at the chosen w^*. Rearrange (8), given θ, we have:

$$-b(\theta) = -\langle \theta, w^*(\theta) \rangle + \psi(w^*(\theta)) \tag{9}$$

Here we have a one-to-one mapping from θ to w^* because we assume ψ is convex and smooth. Otherwise, we cannot write w^* as a function of θ.

5.1.2 Conjugate

Now we discuss what is a conjugate function. Consider a slope line passing through the origin: $\langle \theta, w \rangle$, and the function $\psi(w)$. As illustrated in Fig. 4, we want to look for the maximum distance between the two functions, i.e. $w_\theta^* = \arg\max_w (\langle \theta, w \rangle - \psi(w))$. One observation we can make is

that at w_θ^*, the slope of the tangent line is same as θ. If we draw the tangent line (dotted), the intercept of the tangent line is actually equal to the maximum distance between two functions at w_θ^*.

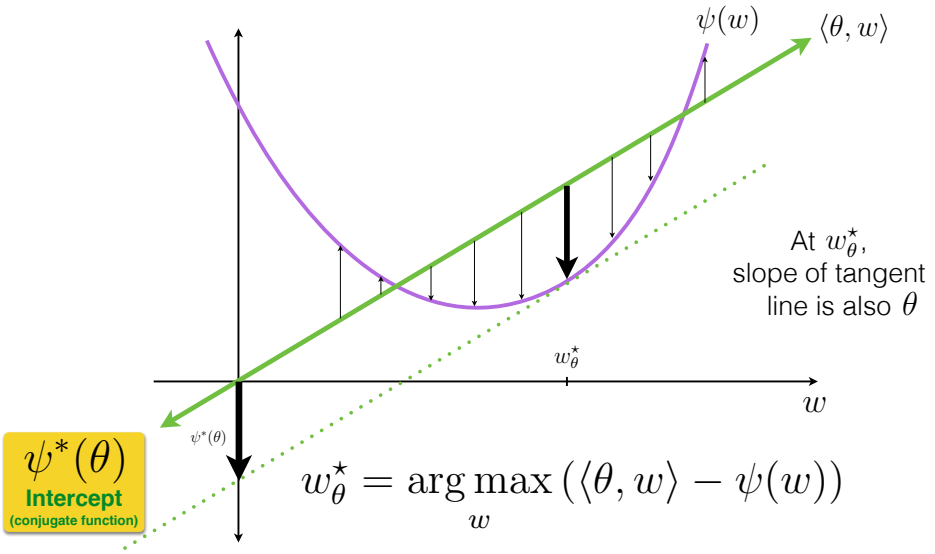

Figure 4: Geometry of the conjugate function $\psi^*(\theta)$

We call this maximum distance the "intercept" $\psi^*(\theta)$ and for any slope θ, we have:

$$\psi^*(\theta) = \max_w (\langle \theta, w \rangle - \psi(w)) \tag{10}$$

This is the geometric interpretation of the conjugate function (4). To reiterate, the convex conjugate at θ is the maximum distance between function ψ and the slope function $\langle \theta, w \rangle$ passing through the origin.

5.2 Proofs for properties of convex conjugate

Proof for (5):

Proof.

$$\begin{aligned}
\psi^*(\boldsymbol{\theta}) &= \max_w (\langle \boldsymbol{\theta}, \boldsymbol{w} \rangle - \psi(\boldsymbol{w})) && \text{Definition (10)} \\
\nabla_{\boldsymbol{\theta}} \psi^*(\boldsymbol{\theta}) &= \nabla_{\boldsymbol{\theta}} \max_w (\langle \boldsymbol{\theta}, \boldsymbol{w} \rangle - \psi(\boldsymbol{w})) && \text{Take derivative} \\
&= \nabla_{\boldsymbol{\theta}} (\langle \boldsymbol{\theta}, \boldsymbol{w}^* \rangle - \psi(\boldsymbol{w}^*)) && \text{In terms of optimal } \boldsymbol{w}^* \\
&= \boldsymbol{w}^* && \text{Compute partial derivative}
\end{aligned}$$

□

5.3 Geometric interpretation of the Bregman divergence

Fig. 5 shows the geometric interpretation of Bregman divergence [3]. $\psi(\boldsymbol{w}) - \psi(\boldsymbol{u})$ is difference of function values. $\nabla\psi(\boldsymbol{u})^\top(\boldsymbol{w}-\boldsymbol{u})$ is the first order approximation of $\psi(\boldsymbol{w})$. $D_\psi(\boldsymbol{w}||\boldsymbol{u})$ is the Bregman divergence. Geometrically, it is the "distance" between the point on the function ψ (regularization function in the case of OMD) and its tangent at $\boldsymbol{u}$ when evaluated at $\boldsymbol{w}$. It represents the first order approximation error evaluated at $\boldsymbol{w}$ with linearization point $\boldsymbol{u}$.

$$D_\psi(\boldsymbol{w}||\boldsymbol{u}) = \psi(\boldsymbol{w}) - \psi(\boldsymbol{u}) - \nabla\psi(\boldsymbol{u})^\top(\boldsymbol{w}-\boldsymbol{u})$$

$$\psi(\boldsymbol{w}) - \psi(\boldsymbol{u}) = \nabla\psi(\boldsymbol{u})^\top(\boldsymbol{w}-\boldsymbol{u}) + D_\psi(\boldsymbol{w}||\boldsymbol{u})$$

Difference of function values — First order approximation of $\psi(\boldsymbol{w})$ — Bregman Divergence (approximation error)

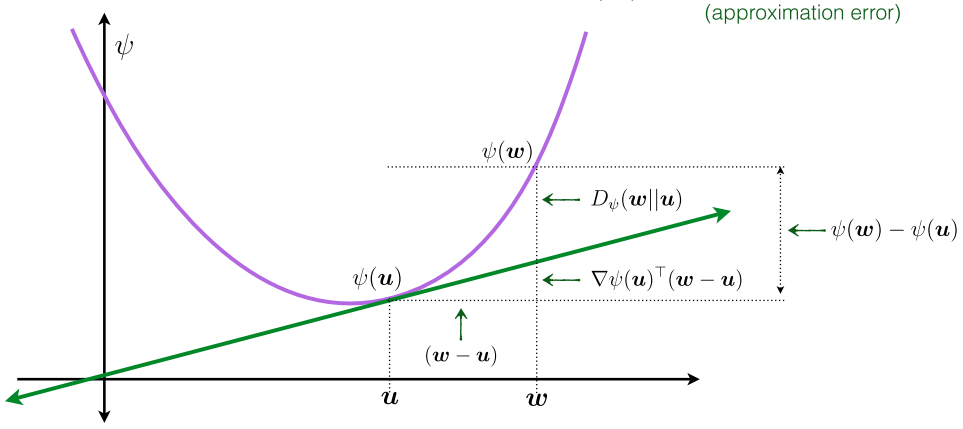

Figure 5: Geometric interpretation of Bregman divergence